Quelques rappels de statistiques

Philippe Veber

17 octobre 2023

Test de significativité

Un exemple concret

L'usine Duclou vient de recevoir une nouvelle machine censée produire des vis de longueur 10 mm avec un écart-type de 0.1 mm. Les ouvri · er · ères décident de la tester avant de la mettre en production, et produisent pour cela 10 vis. La longueur moyenne trouvée dans cet échantillon est de 10.098.

Comment décider si la machine a effectivement les caractéristiques annoncées par le constructeur ?

Des simulations ?

Prenons le constructeur au mot, et supposons que la machine produit effectivement des vis dont la longueur varie aléatoire selon une distribution normale, de moyenne 10 et d'écart type 0.1

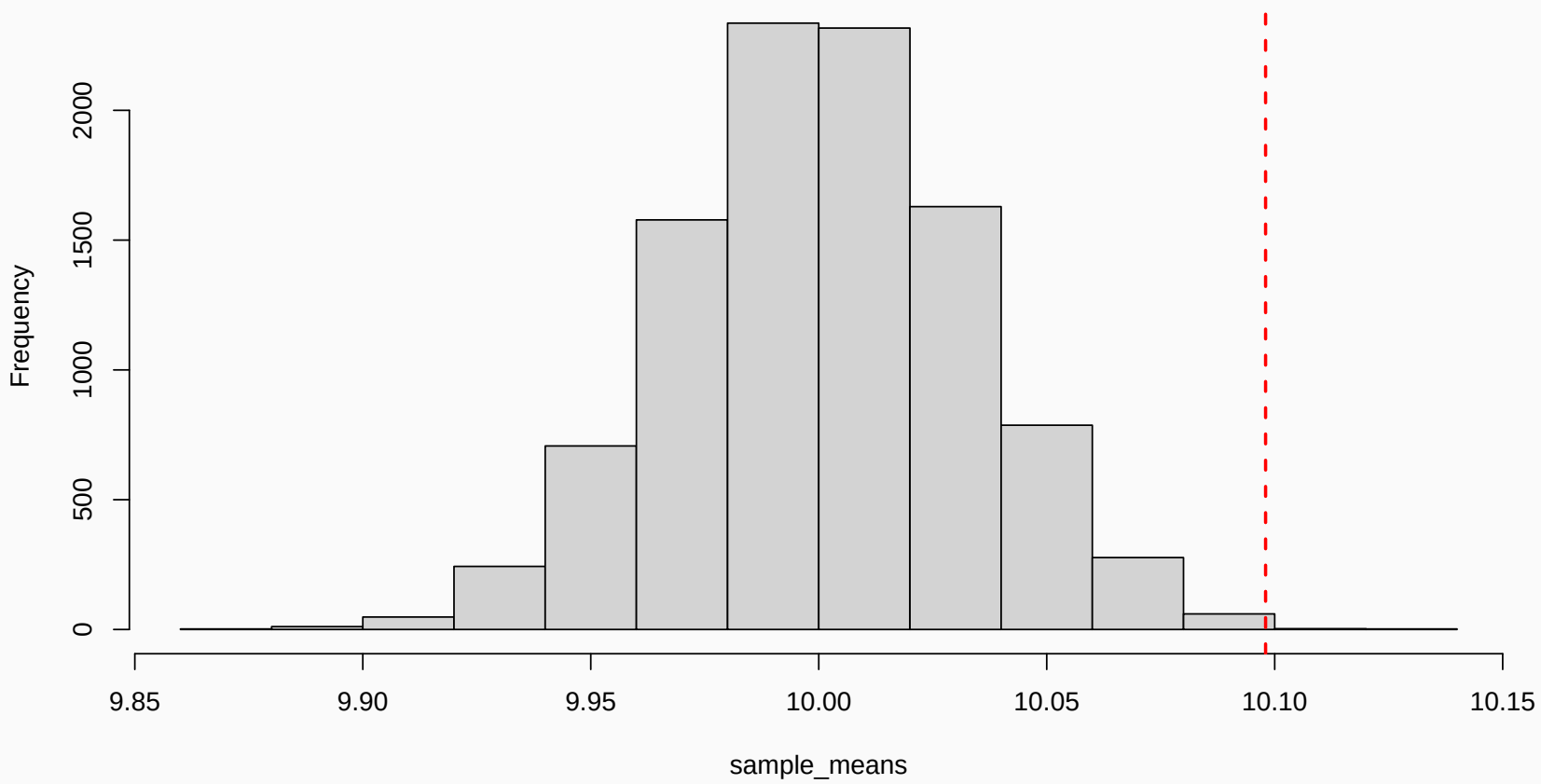
On peut simuler facilement sous R un échantillon de taille 10 et en calculer la moyenne

```
mean(rnorm(10, mean = 10, sd = 0.1))
```

mais cette valeur fluctue d'un échantillon à l'autre.

Distribution de la moyenne d'échantillon

```r
N <- 10000
sample_means <- replicate(N, mean(rnorm(10, mean = 10, sd = 0.1)))
hist(sample_means, main = "")
abline(v = 10.098, col = "red", lwd = 2, lty = 2)
```



Valeur observée vs distribution de la moyenne d'échantillon

On peut compter le nombre de simulations dans lesquelles l'écart à la moyenne attendue est plus grand que celui observé

```
sum(abs(sample_means - 10) > 0.098) / N
```

```
## [1] 0.0021
```

Par conséquent, *si on suppose la machine bien réglée* l'échantillon produit paraît suspicieusement atypique !

La démarche du test de significativité

On suit une sorte de raisonnement par l'absurde. On considère une hypothèse, notée H_0 et appelée **hypothèse nulle**, que l'on cherche à **réfuter** à partir d'observations expérimentales.

1. on collecte des données
2. on choisit un résumé de ces données T, appelé statistique de test, et on calcule sa valeur t_{obs} sur les données observées
3. on détermine la probabilité p, si on répétait infiniment l'expérience dans les conditions décrites par H_0, de trouver une valeur de T supérieure à t_{obs}

p est appelée p-**valeur**

Plus p est faible, plus les données observées sont invraisemblables sous H_0 et plus on sera amené à penser que H_0 est fausse.

Illustration sous R : le test t de Student

```
> lengths <- c(10.01,10.24,10.26,10.19,9.94,10.34,
               10.25,10.14,10.02,9.95)
> t.test(lengths, mu = 10)

    One Sample t-test

data:  lengths
t = 2.9473, df = 9, p-value = 0.01629
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
 10.03115 10.23685
sample estimates:
mean of x
   10.134
```

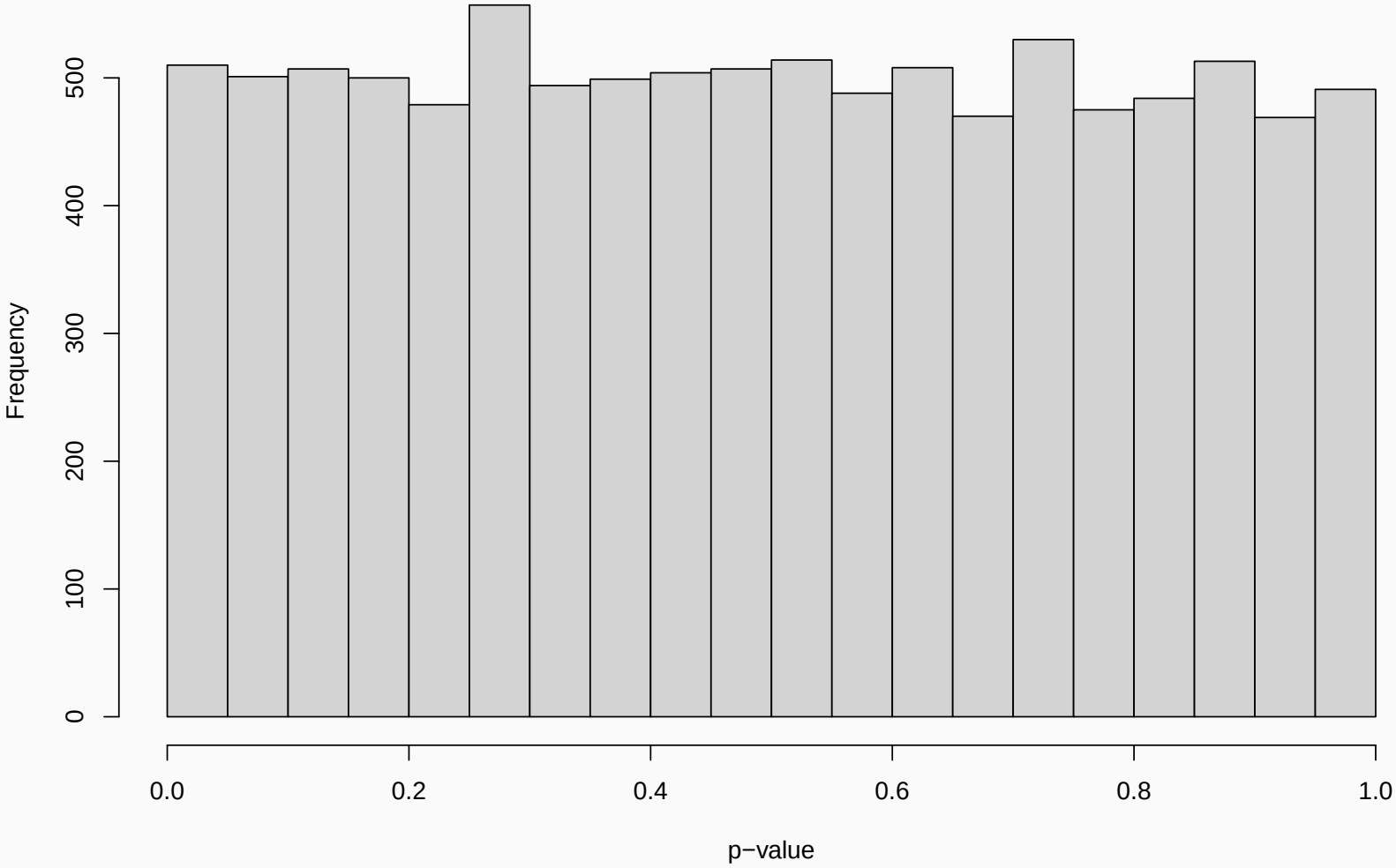
Distribution de la p-valeur sous H_0

On a vu qu'une p-valeur faible suggère que H_0 est fausse. Mais comment se comporte la p-valeur quand H_0 est vraie ?

On peut s'en faire une idée par simulation

```
pvals <- replicate(N, t.test(rnorm(10), mu = 0)$p.value)
hist(pvals, main = "", xlab = "p-value")
```

Distribution de la p-valeur sous H_0



p-valeur et taille d'effet

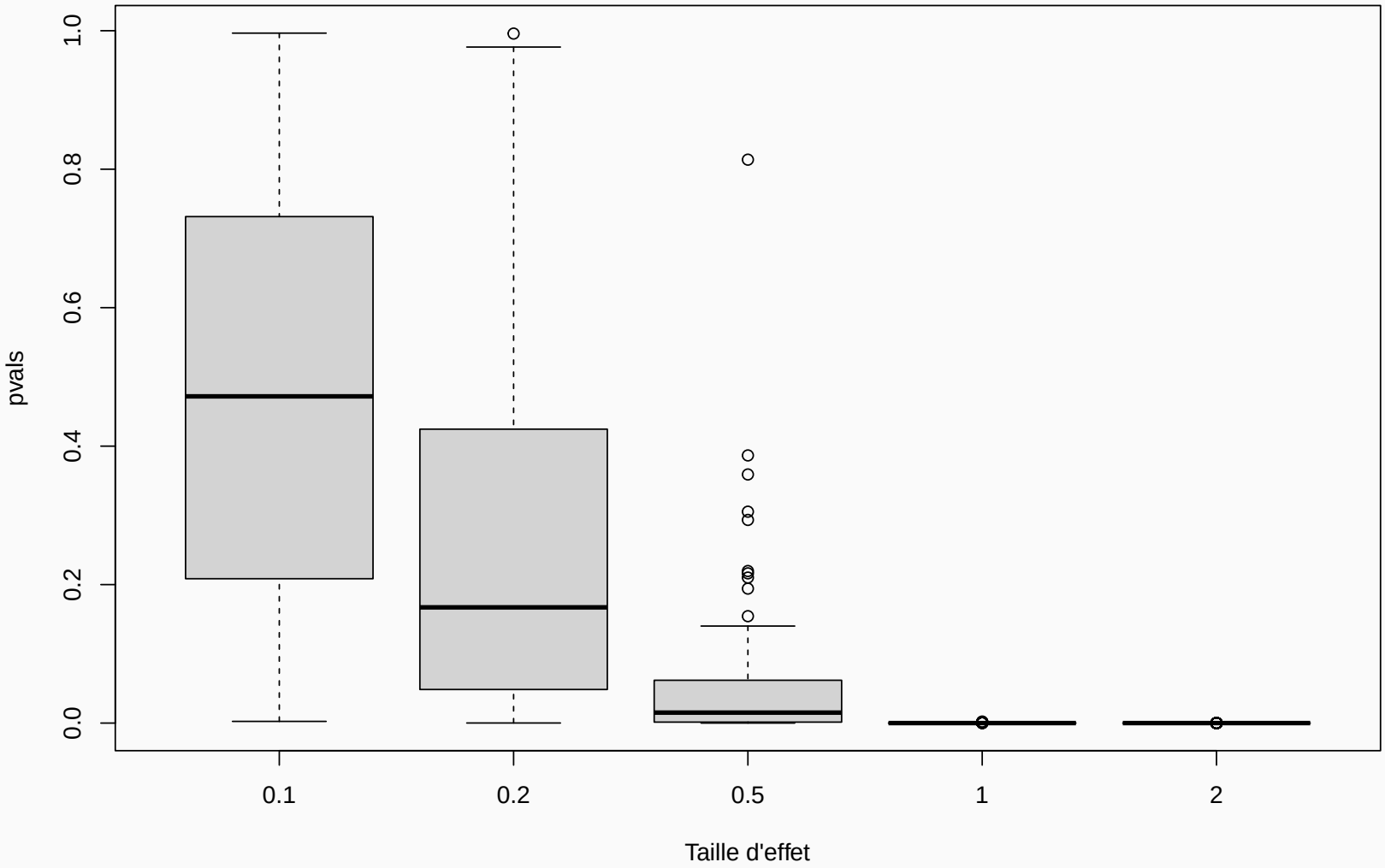
Lorsqu'on compare deux conditions expérimentales, on appelle taille d'effet la variation d'une grandeur d'intérêt entre les deux conditions.

P. ex., en transcriptomique la taille d'effet est le plus souvent le différentiel d'expression d'un gène entre deux conditions

L'hypothèse correspond le plus souvent à une taille d'effet nulle. Mais comment la p-valeur évolue lorsque la taille d'effet augmente ?

Pour l'illustrer, on simule une loi normale $N(\mu, 1)$ où μ est la taille d'effet, et on teste contre l'hypothèse nulle $\mu = 0$

p-valeur et taille d'effet



p-valeur et taille d'échantillon

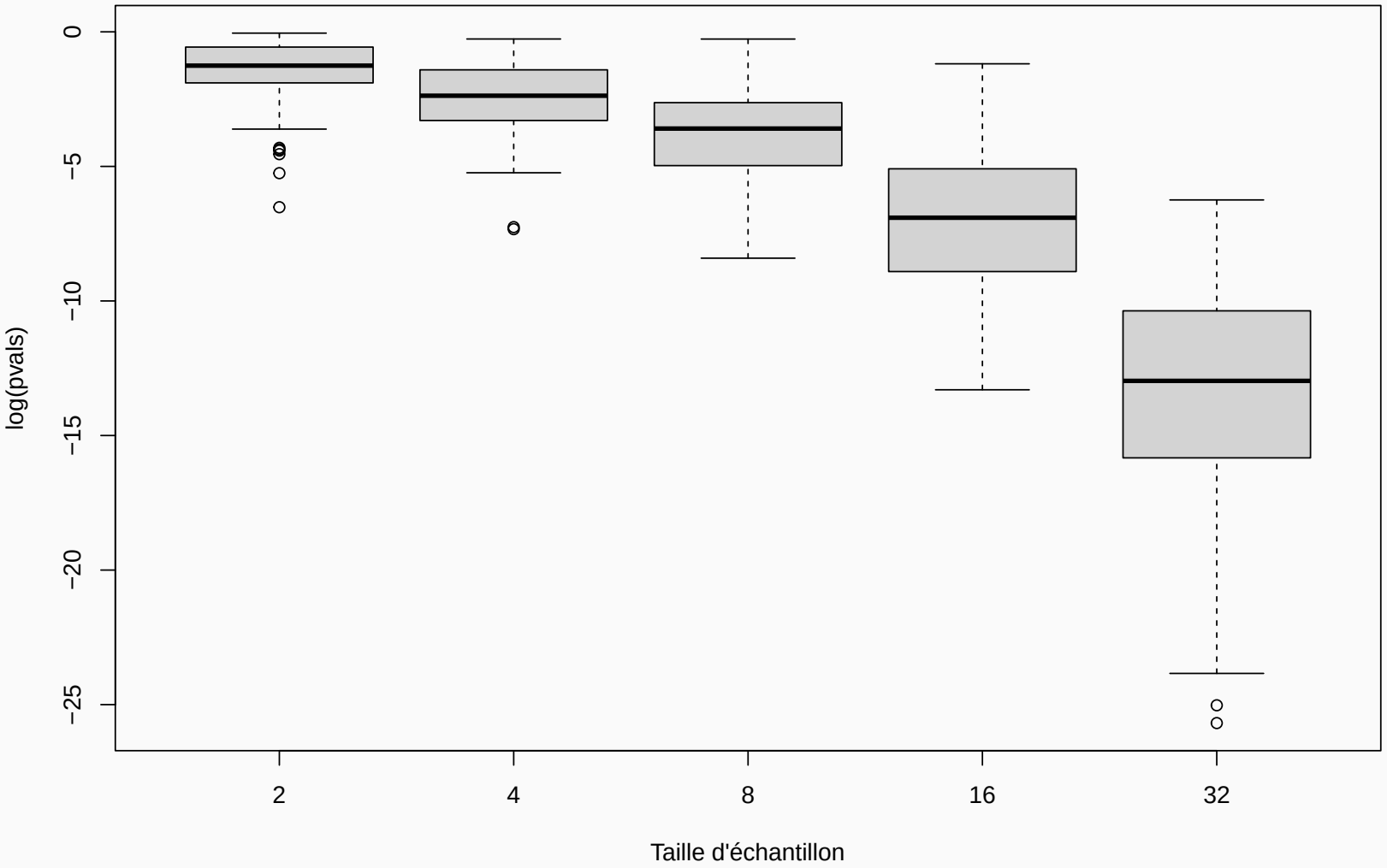
Avec un échantillon plus grand, on peut fournir des estimations plus précises :

- plus proche de la "vraie" valeur
- avec une incertitude moindre

Comment la taille d'échantillon influence-t-elle la p-valeur ?

Pour l'illustrer on simule sous $N(1, 1)$ et on teste contre $\mu = 1$

p-valeur et taille d'échantillon



Exercice

On considère deux populations d'insectes dont on mesure la taille. Est-il possible de détecter une différence de taille moyenne entre les deux populations de 0,01 mm sachant que l'écart-type dans chaque population est de 1 mm ?

Test multiple

Seuil de p-valeur ?

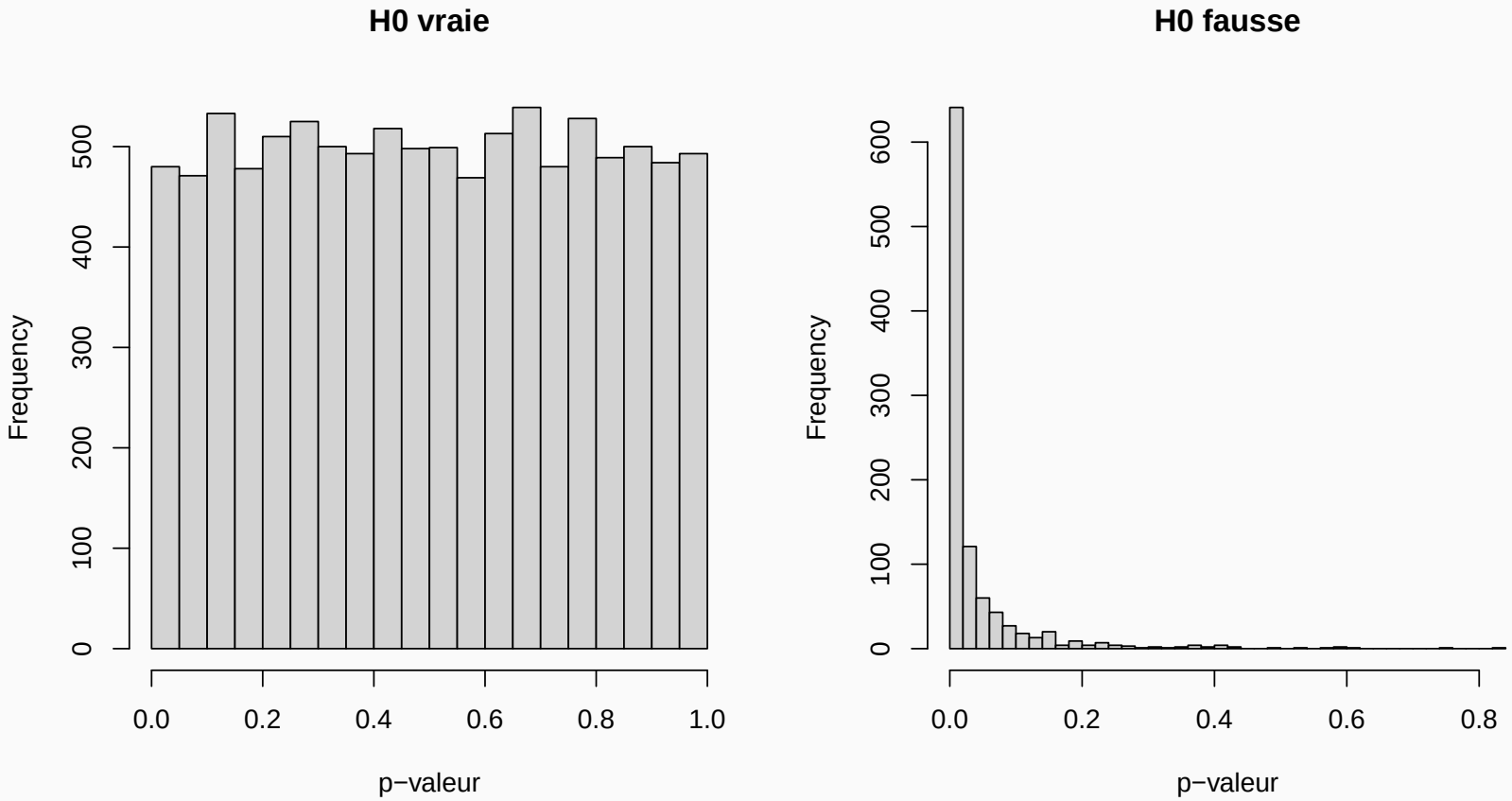
Comment décider, en pratique, si une p-valeur est suffisamment faible pour rejeter l'hypothèse nulle ?

- le seuil à 0.05 est à proscrire (cf *infra*)
- il n'existe pas de seuil universel
- dépend en pratique d'autres considérations (coût d'un faux-positif, d'un faux-négatif)

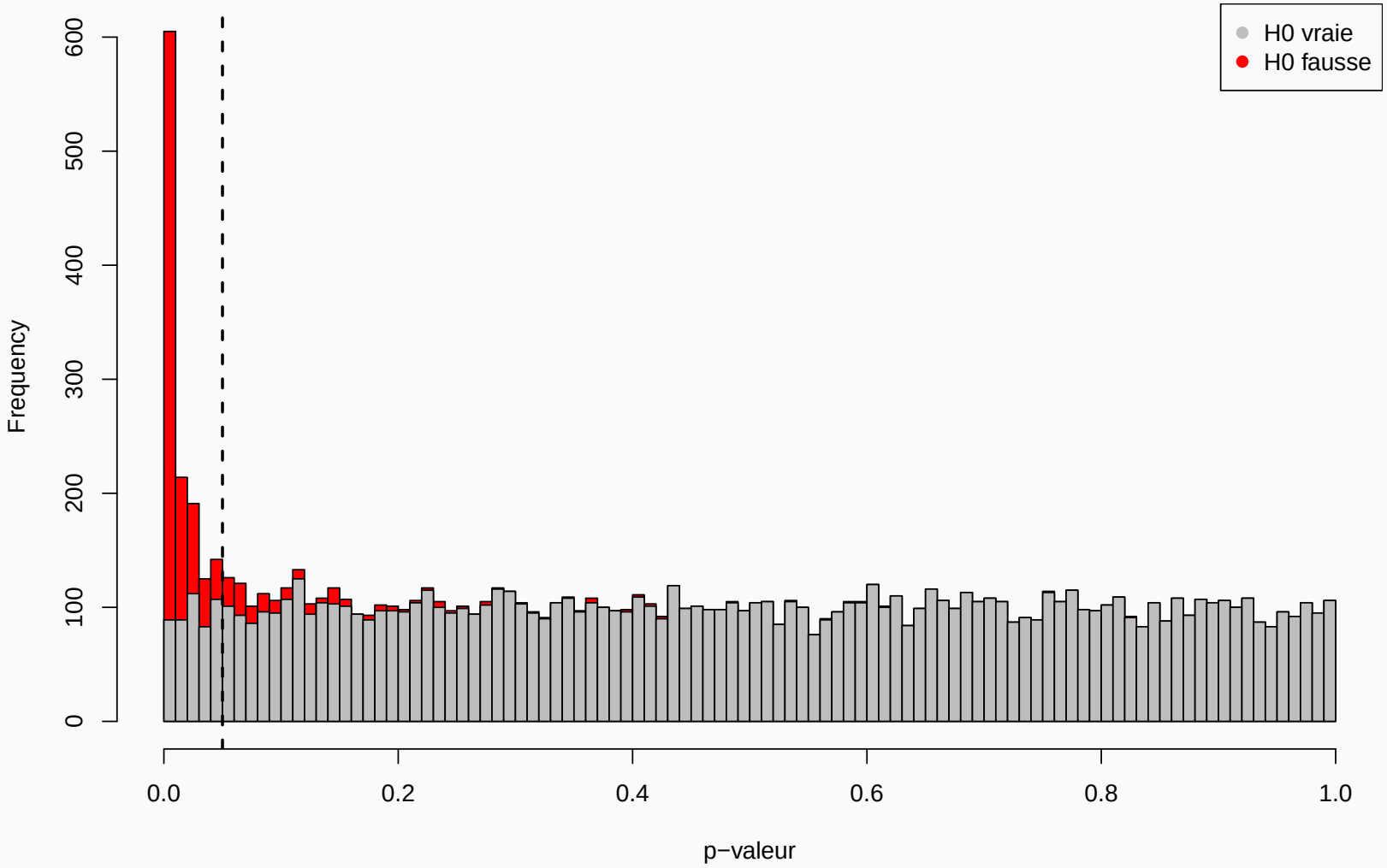
Test multiple

Supposons qu'au lieu d'effectuer un seul test, on en fasse un grand nombre. Pour une partie, H_0 est vraie, pour l'autre elle est fausse.

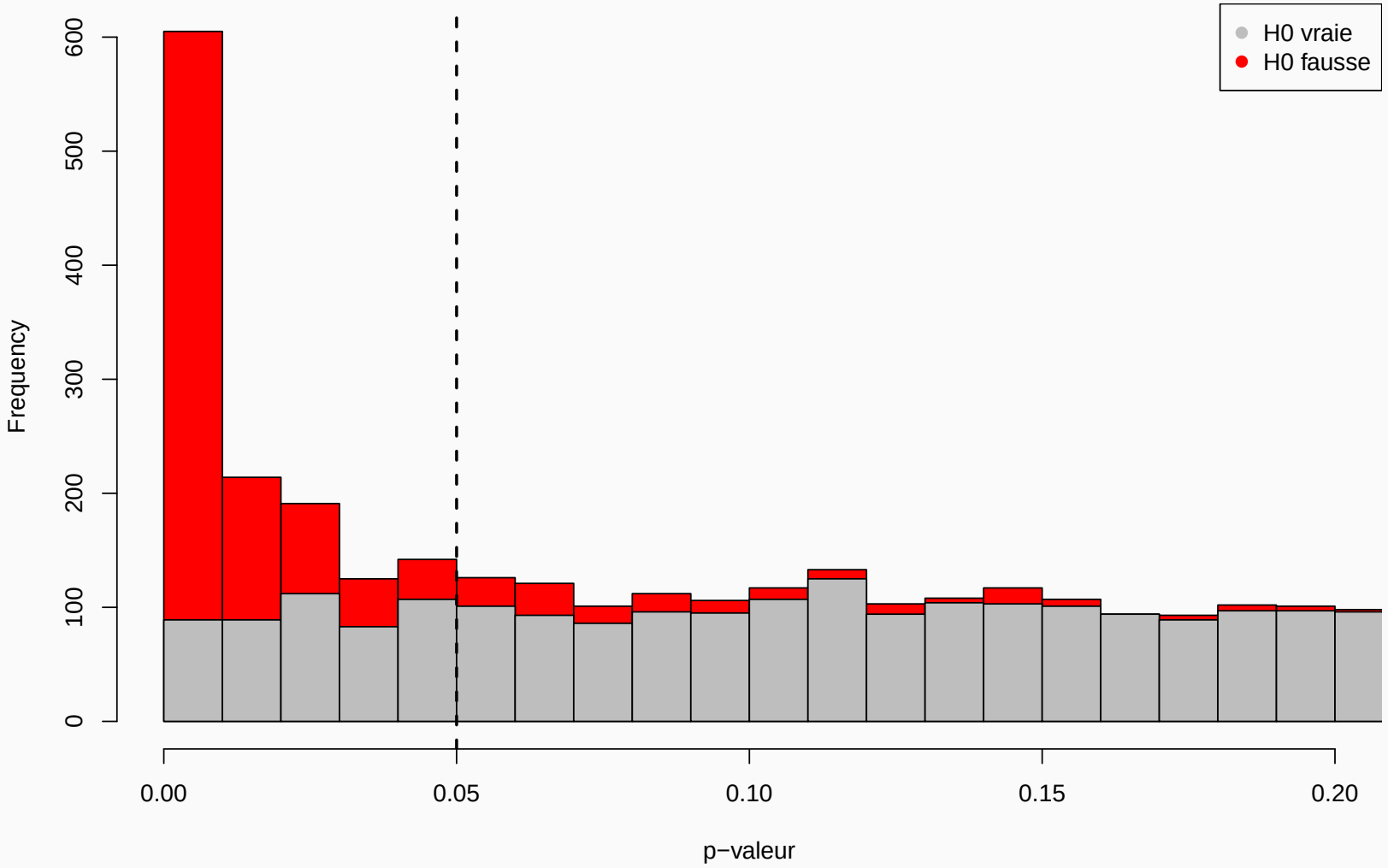
Les p-valeurs sont typiquement distribuées ainsi :



Mélange des distributions de p-valeur



Mélange des distributions de p-valeur



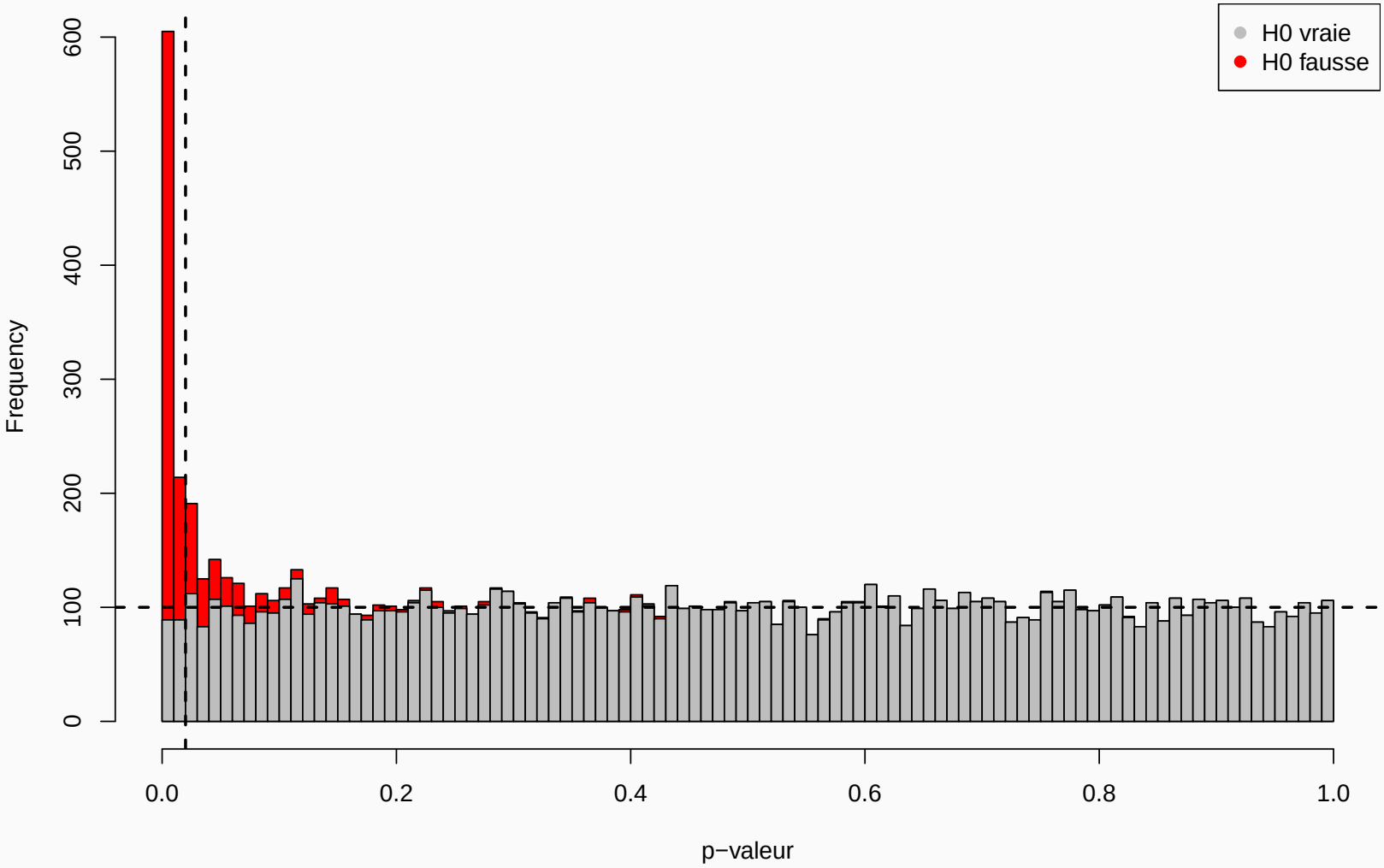
Taux de fausse découverte

- en choisissant un seuil à 0.05, on a (sur cet exemple) environ 39% des hypothèses nulles rejetées qui sont en fait vraies
- entre 0.04 et 0.05 cette proportion monte à 69%
- $\Rightarrow$ on oublie le seuil à 0.05
- la proportion d'hypothèses nulles rejetées qui sont en fait vraies est appelée **taux de fausse découverte** (*False Discovery Rate*)
- elle constitue un critère commode en pratique pour fixer un seuil
  - on sait assez facilement quel taux de "bruit" on peut se permettre
  - on sait l'estimer facilement à partir de l'ensemble des p-valeurs

Correspondance seuil de p-valeur / taux de fausse découverte

seuil p-valeur	FDR	FDR estimée
0.05	0.39	0.41
0.04	0.34	0.38
0.03	0.30	0.32
0.02	0.25	0.26
0.01	0.18	0.18
0.001	0.05	0.06

Détermination géométrique du taux de fausse découverte



Remarques

- le test multiple est souvent présenté comme un “danger”, qui obligerait à “corriger” ou “ajuster” les p-valeurs
- cette vision est incorrecte
  - les p-valeurs ne sont pas “faussées” par le test multiple
  - test multiple ou pas, elles expriment une information difficile à interpréter
  - le test multiple offre l’opportunité de les exploiter pour calculer une grandeur nettement plus utile, le taux de fausse découverte